VIDA: A Visual Intent-driven Design Assistant for Proactive Multimodal Clarification

Yanshan Liu[1]* **Hongbo Zhang**[2]* **Zhen Sun**[1] **Jiaheng Wei**[1] **Kaishun Wu**[1]†
[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]Beijing Wuzi University

Abstract

In complex domains like interior design, user requests are often ambiguous and multimodal. Professional designers address this by asking strategic clarification questions based on hierarchical priorities, a capability lacking in current Vision-Language Models (VLMs). When fine-tuned on dialogue data, existing models often exhibit modality forgetting, overfitting to textual patterns while neglecting visual cues and thus producing hallucinated or visually irrelevant questions. To bridge this gap, we introduce **VIDA**[1] (Visual Intent-driven Design Assistant), an assistant designed to generate proactive, visually grounded, and strategically prioritized clarification questions. Instead of standard fine-tuning, we propose a strategy-aware alignment framework that evolves from imitation learning to value-driven reinforcement. We utilize Group Sequence Policy Optimization to strictly enforce expert protocols, ensuring the model not only mimics fluent speech but also adheres to optimal inquiry strategies. Crucially, we design a novel hierarchical reward mechanism with Dynamic Intent Binding to align the assistant with professional prioritization standards. To facilitate this research, we construct and release **InteriorClarify**, a multimodal benchmark dataset comprising 1,016 real-world consultation cases annotated with this three-tier intent hierarchy. Extensive experiments demonstrate that **VIDA** sets a new state-of-the-art, improving the Strategic Alignment Score (SAS) by 20.59% over SFT baselines and effectively restoring visual grounding capabilities lost during standard fine-tuning.

1 Introduction

Recent advances in large language models (LLMs) and vision-language models (VLMs) have demonstrated remarkable capabilities in text generation and instruction following (Brown et al., 2020; Ouyang et al., 2022), continuously pushing the frontier in visual question answering, multimodal understanding, and generation tasks (Alayrac et al., 2022; Li et al., 2022; Liu et al., 2023; Sun et al., 2025; Li et al., 2025). However, many LLM/VLM-based applications implicitly assume that user specifications are complete and unambiguous (Zamfirescu-Pereira et al., 2023; Jian et al., 2025; Kobalczyk et al., 2025; Zhang et al., 2025b). This assumption breaks in open-ended design tasks,

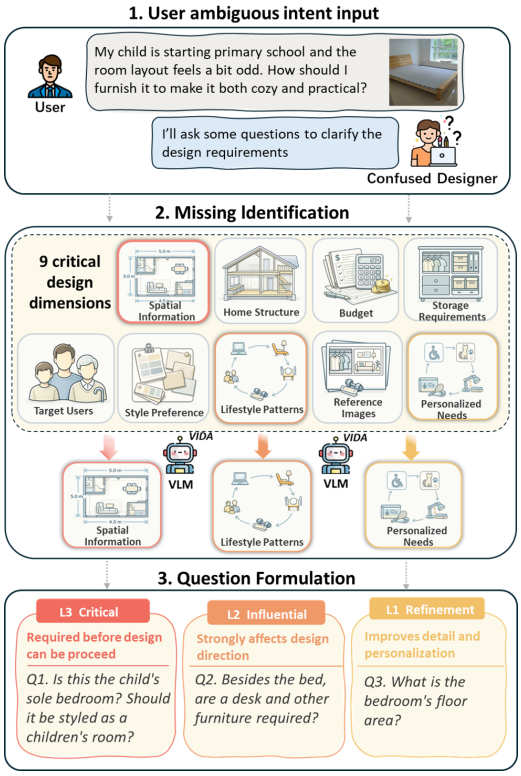

Figure 1: **VIDA** system overview. **VIDA** operates in three stages: (1) the user provides an ambiguous design intent with text and images; (2) a VLM identifies missing elements across nine critical design dimensions; (3) **VIDA** formulates a targeted clarification question by adhering to a three-tier constraint hierarchy.

*Equal contribution.
†Corresponding author: wuks@hkust-gz.edu.cn
[1]The data and code are publicly available at https://github.com/Loewen-Hob/VIDA.git.

Findings of the Association for Computational Linguistics: ACL 2026, pages 33793–33804
July 2-7, 2026

where requests are often underspecified and effective systems must refine intent through interaction.

Interior design consultation exemplifies this setting. Underspecification arises from (i) conceptual ambiguity, where the same descriptor can be realized through different choices (e.g., lighting temperature, materials, textures, or color palette) and varies across users and contexts (Mehrabi et al., 2023), and (ii) information incompleteness, where specifications omit constraints needed for implementable solutions (e.g., spatial feasibility, functional requirements, and budget). Professional designers rarely implement such requests directly. Instead, they conduct clarification dialogues that prioritize what must be decided early (e.g., layout) and defer refinements (e.g., decorative details). This practice suggests a key requirement for effective assistants: beyond asking relevant clarification questions, they must select and sequence questions across turns to efficiently reduce uncertainty under limited time and turn budgets.

This requirement is naturally related to Clarification Question Generation (CQG), which has been studied in dialogue systems and information retrieval (Rao and Daume III, 2018; Aliannejadi et al., 2021; Xu et al., 2019; Kumar and Black, 2020). Yet existing CQG methods face three fundamental obstacles in professional design settings. First, most methods are text-only (Xu et al., 2019; Mu et al., 2024), while design consultations rely heavily on visual evidence such as room photos and reference images; existing multimodal systems are rarely optimized for missing-constraint identification (Ramezan et al., 2025). Second, existing CQG approaches rarely model the strategic ordering of questions across turns, despite that design consultations involve unique spatial conditions and personalized preferences that cannot be exhaustively predefined. Current retrieval methods (Rao and Daume III, 2018; Aliannejadi et al., 2021) suffer from limited candidate coverage, while generative approaches lack explicit priority guidance. This neglect results in suboptimal information convergence, where systems fail to address critical constraints efficiently. Third, prior work focuses on single-turn utility instead of sequence-level strategy, even though question importance and optimal ordering depend on context in real consultations.

To address these limitations, we propose **VIDA** (***Visual Intent-driven Design Assistant***), a multimodal clarification system that explicitly models constraint priority across turns. As illustrated in Figure 1, **VIDA** is trained with reinforcement learning and hierarchical reward shaping to categorize constraints into three tiers: critical, influential, and refinement. This design encourages the system to elicit high-impact constraints early while remaining turn-efficient. By jointly processing visual and textual inputs, **VIDA** grounds clarifications in the actual spatial context.

Progress on this problem is also constrained by data: existing CQG resources rarely capture priority-driven, multimodal clarification behavior in realistic interior design consultations. We therefore introduce **InteriorClarify**, a multimodal benchmark dataset of 1,016 real interior design consultation sessions annotated by professionals with tiered constraint labels, enabling systematic training and evaluation of priority-aware clarification strategies in realistic design scenarios.

Our main contributions are as follows:

- We formalize a professionally grounded three-tier constraint hierarchy and introduce **VIDA**, which leverages VLMs and hierarchical RL to select and order clarification questions under turn budgets.
- We release **InteriorClarify**, a multimodal dataset of 1,016 interior design consultation sessions annotated by professional designers.
- Extensive evaluations show that **VIDA** sets a new state-of-the-art in multimodal clarification. It surpasses strong baselines, securing the highest scores in both strategic alignment (49.02% SAS) and intent accuracy (82.35% K-RME).
- We further conduct a blind expert preference study to evaluate the practical utility of VIDA in realistic interior design consultation settings.

2 Related Work

2.1 Clarification Question Generation

Text-based Approaches. Early CQG studies modeled clarification as ranking/selection to maximize expected information gain (Rao and Daume III, 2018), and later moved to generative formulations that directly produce questions (Rao and Daumé, 2019). This line was extended to open-domain information-seeking and conversational search settings (Aliannejadi et al., 2019, 2021; Zamani et al., 2020), as well as domain- or scenario-specific applications such as knowledge-based QA and technical community forums (Xu et al., 2019; Kumar and Black, 2020). More recently, task-oriented CQG

has emphasized task knowledge and user personalization (Feng et al., 2023), with reinforcement learning and preference-based optimization used to improve multi-turn dialogue policies (Zhang et al., 2025a). Despite progress, most text-based CQG remains single-modality and often evaluates questions in isolation, limiting systematic constraint elicitation when visual evidence is crucial in professional consultations.

Multimodal Approaches. Recent work has explored multimodal CQG by incorporating visual signals to better infer user intent. In conversational search, adding relevant images improves single-turn retrieval and has been extended to multi-turn text, like image interactions (Yuan et al., 2024; Ramezan et al., 2025). Related efforts study interactive disambiguation for ambiguous visual questions and benchmarks for evaluating VLM disambiguation (Jian et al., 2025), as well as spatial-reasoning-driven clarification in embodied AI and collaborative visual dialogue (Shi et al., 2022, 2023; Madureira and Schlangen, 2023). Despite incorporating vision, most prior work is limited to open-domain search or generic VQA with synthetic dialogues, and may still suffer from visual unreliability, failing to model domain-specific constraints and priority-aware acquisition grounded in professional consultations (Sun et al., 2026).

2.2 Reinforcement Learning for Dialogue Systems

Reinforcement learning (RL) is widely used for dialogue policy learning by casting multi-turn interaction as a sequential decision process optimized via cumulative rewards, enabling strategic decisions such as when to ask, confirm, or respond. In LLMs, RL from Human Feedback and its variants align outputs with human preferences through reward-based policy optimization (Ouyang et al., 2022), while Direct Preference Optimization offers a simpler alternative (Rafailov et al., 2023).

In dialogue systems, RL has been applied to task-oriented policy optimization and reward learning for question generation, including clarification timing and formulation (Kwan et al., 2023; Pan et al., 2019; Hu et al., 2020). Recent work has increasingly incorporated RL into LLM-based dialogue systems to improve multi-turn decision-making (Guan et al., 2023; Zhang et al., 2025a; Wang et al., 2025), such as through enhanced semantic representations and explicit modeling of future turns. In interior design, RL has also been explored for interactive recommendation with multi-turn user feedback (Zhang et al., 2023). However, most RL-based dialogue work targets generic settings and does not explicitly model domain constraints and their priorities in professional consultations, where specifications are incomplete and efficient interaction requires proactively eliciting decision-critical information.

3 The InteriorClarify Benchmark Dataset

3.1 Data Collection and Taxonomy

We introduce **InteriorClarify**, a novel multimodal clarification question benchmark dataset comprising $1,016$ interior design consultation cases, where each case pairs an ambiguous client request with an accompanying image and includes the clarification questions provided by designers.

The data sources comprise (i) online community platforms: REDNote, Zhihu, Douban, and Tieba [2] ($\approx$ 85%); (ii) online consultation chat records ($\approx$ 5%); and (iii) offline consultation recordings ($\approx$ 10%). Online chat records and offline recordings contain entire real consultation dialogues, including clients' ambiguous requests and designers' clarification questions elicited in situ. By contrast, online community platforms mostly provide standalone renovation requests (posts, comments, and group chats) without designer follow-ups.

To address this, we invited 10 professional interior designers to write clarification questions based on their consultation experience, thereby standardizing the data format. As illustrated in Figure 2, the data collection and annotation process comprises three data sources and a systematic three-tier annotation workflow. Among the $1,016$ cases, 821 contain photographs of clients' existing living spaces, while the remaining 195 are text-only and contain no visual input. To standardize the multimodal input format, we generated accompanying images for these 195 instances using Qwen-image (Wu et al., 2025) based on the corresponding text descriptions. Client initial requests average 28 Chinese characters in length, reflecting the highly concise and ambiguous nature of real design consultations. All data were originally collected in Chinese and were translated into English via a pipeline utilizing Qwen2.5-7B-Instruct and Baidu Translate.

To systematically capture information required

[2]These are major Chinese online platforms where users commonly seek and share advice on home decoration.

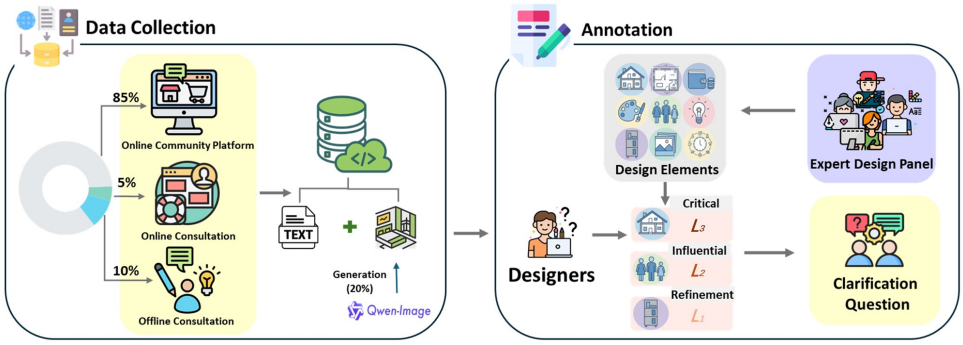

Figure 2: Data collection and annotation pipeline. Left: data sources and image completion for text-only cases. Right: 10 professional designers annotate each case by identifying missing elements across nine design dimensions, assigning them to a three-tier importance hierarchy, and writing the corresponding clarification questions.

in interior design consultations, we worked with experienced designers through multiple rounds of discussion to develop a taxonomy of nine critical design elements (see Table 1). These elements span four aspects that support turning ambiguous requests into actionable plans: structural constraints, functional requirements, aesthetic preferences, and personalization needs.

Building on this taxonomy, we further propose a three-tier importance framework (L_3 / L_2 / L_1) to characterize the priority of different design elements in the design process. L_3 (Critical Information) represents necessary prerequisites for design work; missing this information prevents designers from initiating any substantive planning. L_2 (Influential Information) encompasses elements that significantly impact core design directions; their absence introduces substantial uncertainty in spatial organization or aesthetic positioning. L_1 (Refinement Information) is used to optimize design details and enhance personalization; its absence does not affect the basic feasibility of the plan but reduces design refinement.

Based on this framework, the clarification process in the dataset is organized into up to three sequential rounds: the first round targets missing L_3 elements, the second round addresses L_2 elements, and the third round focuses on L_1 elements, thereby efficiently acquiring critical information for design within limited interactions.

3.2 Annotation Process

The dataset was annotated by ten professional interior designers with practical consultation experience.(see Figure 2, right panel) All annotations followed unified annotation guidelines that clearly specify the definitions of the nine design elements, criteria for determining the three importance levels (L_3/L_2/L_1), and principles for generating clarification questions. All designers received centralized training before formal annotation.

Prior to formal annotation, each designer completed a small-scale pilot annotation (10 samples per designer) to calibrate annotation standards. During this phase, we identified that "Spatial Information" and "Home Structure" exhibited boundary ambiguity in certain cases. Through collective discussion with designers and supplementary clarification of standards, we unified the relevant judgment rules before formal annotation began.

The annotation process followed a uniform three-step procedure. For each client request, designers first identify missing L_3 (Critical Information) design elements and generate corresponding clarification questions for each missing element. They then process L_2 (Influential Information) and L_1 (Refinement Information) levels in the same manner. Each instance produces up to three clarification questions, corresponding to the three levels of importance.

Upon completion, two authors conducted a systematic review of all 1,016 instances, focusing on the accuracy of missing element identification, appropriateness of importance classification, and the professionalism and consistency of clarification questions. For a small number of ambiguous cases, authors engaged in retrospective discussion with the corresponding designers to reach consensus. The entire annotation process was completed between November and December 2025.

3.3 Dataset Statistics

The dataset comprises 1,016 interior design consultation cases spanning four primary room types:

Element	Description
Home Structure	Structural and hard decoration aspects including walls, ceilings, and fixed installations
Spatial Information	Space dimensions, size constraints, and functional requirements of rooms
Budget	Financial constraints and cost expectations for the design project
Target Users	Occupants of the space
Style Preference	Aesthetic directions and visual themes
Lifestyle Patterns	Daily routines, activity flows, and living habits
Personalized Needs	Unique client requirements, hobbies, and special accommodations
Storage Requirements	Needs for cabinets, shelving, and organizational solutions
Reference Images	Reference visuals conveying desired aesthetics or specific features

Table 1: Design Element Taxonomy.

Design Element	L_3	L_2	L_1	Total
Home Structure	65	32	19	116
Spatial Information	325	62	27	414
Budget	21	123	194	338
Target Users	24	50	44	118
Style Preference	361	211	76	648
Lifestyle Patterns	69	120	66	255
Personalized Needs	59	235	203	497
Storage Requirements	40	91	111	242
Reference Images	52	90	264	406
Total	**1,016**	**1,014**	**1,004**	**3,034**

Table 2: Distribution of design elements across importance tiers.

living rooms (496 cases, 48.8%), bedrooms (311 cases, 30.6%), kitchens (101 cases, 9.9%), and bathrooms (108 cases, 10.6%). This distribution approximately reflects the frequency distribution of renovation inquiries for different rooms in real-world design consultations.

Table 2 presents the occurrence frequency of the nine design elements across the three importance levels. As shown, structurally critical elements such as Home Structure and Spatial Information appear frequently at the L_3 level, while refinement-oriented elements such as Budget and Reference Images are primarily concentrated at the L_1 level. Style Preference is the most frequently occurring element with 648 total occurrences, indicating its central role in design consultations.

Different room types exhibit distinct information priority patterns during the clarification process. In living room consultations, style preference emerges as the primary L_3 concern, appearing in 45.3% of cases, reflecting the visual centrality of this space in home design. Bathrooms and kitchens, by contrast, prioritize spatial information at the L_3 stage (41.3% and 46.5% respectively), indicating that functional constraints dominate their design process. At the L_2 level, bathrooms show heightened attention to budget factors (26.6%), while bedrooms emphasize personalized needs (21.5%) and lifestyle patterns (11.3%). Overall, these differences indicate that room type strongly shapes information priorities and motivates room-specific clarification strategies.

Regarding clarification rounds, among the 1,016 cases, 1,004 (98.8%) involve clarification across all three levels (L_3, L_2, and L_1), 10 (1.0%) involve two levels (L_3 and L_2), and 2 (0.2%) involve only the L_3 level. This pattern indicates that most client requests exhibit multi-level information ambiguity requiring systematic clarification.

Client initial requests average 28 Chinese characters in length. Clarification questions average 32 characters, with L_3-level questions averaging 34 characters, L_2 averaging 33 characters, and L_1 averaging 30 characters, reflecting the overall conciseness yet targeted nature of professional design consultation questions.

3.4 Evaluation Metrics

To rigorously assess the assistant's performance across strategic, accuracy, and visual dimensions, we adhere to a multi-dimensional protocol:

Strategic Alignment Score (SAS). SAS is our primary metric for evaluating hierarchical reasoning capabilities. It measures the rate at which the model successfully targets the *highest-priority* missing intent h^* (where $L_3 \succ L_2 \succ L_1$) identified by experts. For a dataset of size N, SAS is defined as:

$$\text{SAS} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1} \left(\text{Sim}(\hat{q}_i, h_i^*) > \tau \right). \quad (1)$$

where $\mathbb{1}(\cdot)$ is the indicator function and $\hat{q}_i$ is the generated question. A higher SAS indicates adherence to the professional design protocol.

Keyword Recall of Missing Elements (K-RME). While SAS focuses on priority, K-RME evaluates the *broad coverage* of intent identification. It calculates the percentage of generated questions that hit *any* valid missing element ($L_1 \cup L_2 \cup L_3$) in the

ground truth, regardless of hierarchy. This serves as a baseline accuracy metric for intent retrieval.

Visual Grounding Score (V-G). To penalize visual hallucinations, we compute the semantic similarity between the generated question and the ground-truth visual prompt (which describes the scene's visual details). This metric ensures the assistant's inquiries remain grounded in the actual visual context of the room.

Semantic Similarity (Max-Sim). We utilize Sentence-BERT to compute the maximum cosine similarity between the generated question and the set of expert reference questions. This metric assesses the semantic fluency and professional tone of the generated text compared to human designers.

3.5 Broader Impacts

InteriorClarify enriches the ecosystem of general-domain Clarification Question Generation (CQG) datasets. Existing large-scale CQG datasets such as ClarQ (Kumar and Black, 2020) cover 173 StackExchange domains, but typically lack systematic modeling of professional consultation scenarios, particularly regarding information priority hierarchies and acquisition strategies.

InteriorClarify addresses this gap by introducing a hierarchical priority annotation framework (L_3: Critical, L_2: Influential, L_1: Refinement) for interior design consultation. This framework characterizes how professionals strategically order information acquisition under turn constraints, providing a reference methodology for other professional consultation domains seeking to incorporate explicit priority modeling in CQG research.

Beyond research, **InteriorClarify** supports development of AI-assisted design tools. Professional design consultation remains inaccessible to many due to cost and geographic barriers. The 1,016 annotated consultation cases enable AI systems to learn professional clarification patterns, extending design guidance to underserved populations lacking access to expert consultation.

4 VIDA: Visual Intent-driven Design Assistant

In this section, we present **VIDA**, a visual intent-driven design assistant. We first formalize the clarification generation task, then describe our two-stage training framework from supervised fine-tuning to reinforcement learning, and finally introduce the strategic reward mechanism that enforces hierarchical design protocols.

4.1 Task Formulation

Let I denote the visual context and R denote the user's initial request. The objective is to generate a clarification question Q that resolves the ambiguity in the user's intent. We define a hierarchical intent structure $\mathcal{H} = \{L_3, L_2, L_1\}$, representing Style, Layout, and Specific Items, respectively. The priority order is defined as $L_3 \succ L_2 \succ L_1$. Formally, the task is to learn a policy $\pi_\theta(Q|I, R)$ that maximizes the recovery of the highest-priority missing intent $h^* \in \mathcal{H}$ in the current context.

4.2 Two-stage Training Framework

To equip the Assistant with both conversational fluency and strategic reasoning capabilities, we employ a two-stage training pipeline.

Stage 1: Supervised Fine-Tuning (SFT). We first fine-tune a pre-trained VLM (e.g., Qwen-VL) on a curated dataset of (I, R, Q_{ref}) triplets. This stage, treating the task as behavior cloning, enables the model to follow instructions and understand visual content. The objective is the standard autoregressive language modeling loss:

$$\mathcal{L}_{SFT} = -\sum_{t=1}^{T} \log \pi_\theta(q_t|I, R, q_{<t}). \tag{2}$$

Stage 2: Reinforcement Learning via GSPO. While SFT provides basic capabilities, it fails to capture the subtle priority logic ($L_3 \succ L_2$). To bridge this gap, we adopt GSPO (Shao et al., 2024), a policy-gradient method that optimizes relative rewards within output groups, removing the need for a separate critic and reducing memory overhead.

Specifically, for each input $x = (I, R)$, we sample a group of G outputs $\{Q_1, \ldots, Q_G\}$ from the current policy π_θ. We calculate the advantage A_i for each output based on its reward r_i:

$$A_i = \frac{r_i - \mu_r}{\sigma_r + \epsilon}. \tag{3}$$

where μ_r and σ_r are the mean and standard deviation of rewards within the group. The optimization objective is:

$$\mathcal{L}_{GSPO} = \frac{1}{G}\sum_{i=1}^{G}\left(\min\left(\frac{\pi_\theta(Q_i|x)}{\pi_{\text{old}}(Q_i|x)}A_i, \text{clip}(\ldots)A_i\right) - \beta\mathbb{D}_{KL}\right). \tag{4}$$

Model	Type	K-RME (↑)	SAS (↑)	Max-Sim (↑)	V-G (↑)
General Baselines					
LLaVA-1.5-7B	VLM	30.39%	20.59%	0.4534	0.4173
InternVL_3.5-8B	VLM	43.14%	23.53%	0.4740	0.4458
InternVL_3-38B	VLM	40.20%	21.57%	0.4623	0.4349
Qwen2.5-VL-7B	VLM	48.04%	24.51%	0.4615	0.4520
DeepSeek-VL-7B	VLM	50.98%	31.37%	0.4666	0.4775
Qwen3-VL-8B	VLM	59.80%	28.43%	0.4598	0.4406
Llama-3-8B	LLM	49.02%	31.37%	0.4486	0.4082
DeepSeek-V3	LLM	73.53%	42.16%	**0.4741**	0.4572
Ablation Baselines					
Qwen3-VL-8B + SFT	SFT	54.90%	28.43%	0.4737	0.2970
VIDA (Ours)	RL	**82.35%**	**49.02%**	0.4641	**0.5035**

Table 3: **Main Results on Clarification Generation.** The top three results are highlighted in distinct colors: **1st** , 2nd , and 3rd . **K-RME**: Keyword Recall. **SAS**: Strategic Alignment Score. **Max-Sim**: Semantic Similarity. **V-G**: Visual Grounding Score.

This approach encourages the model to generate questions with higher strategic value (higher rewards) compared to the group average.

4.3 Strategic Reward Mechanism

The effectiveness of GSPO relies heavily on the quality of the reward signal. We design a dense reward function R_{total} composed of four dimensions:

$$R_{total} = \lambda_1 R_{Hier} + \lambda_2 R_{Sim} + \lambda_3 R_{Vis} + \lambda_4 R_{Fmt}. \quad (5)$$

Hierarchical Strategy Reward (R_{Hier}). Unlike static slot-filling, design consultation requires dynamic prioritization. We propose **Dynamic Intent Binding** mechanism. For an instance with missing intents $\mathcal{M}$, the target intent t is determined by:

$$t = \arg\max_{h \in \mathcal{M}} \text{Priority}(h). \quad (6)$$

The model receives a positive reward only if it addresses the intent t. This forces the policy to learn the *concept of priority* rather than memorizing specific keywords.

Semantic Consistency Reward (R_{Sim}). To mitigate reward hacking, we measure cosine similarity between the generated question and expert questions with a pretrained Sentence-BERT model.

Visual Grounding Reward (R_{Vis}). To mitigate visual hallucinations, we reward questions that explicitly reference visible elements in the image. We extract visual concepts V from the image caption and calculate the keyword overlap rate.

Format Constraint (R_{Fmt}). We impose penalties for excessive length or non-interrogative formats to ensure the professional conciseness of the assistant.

5 Experiments

5.1 Experimental Setup

Dataset. We constructed a high-quality multimodal instruction tuning dataset derived from real-world interior design consultations. The dataset consists of (I, R, Q) triplets, where I is the room image, R is the user request, and Q is the expert clarification question. The training set contains 914 samples, while 102 samples are reserved for the held-out test set. We standardized the missing intent labels into 9 categories mapped to the hierarchical structure $\mathcal{H}$.

Implementation Details. Our **VIDA** assistant is initialized with Qwen3-VL-8B-Instruct. We first perform SFT for 2 epochs using a learning rate of 1e-4. In the RL stage, we employ GSPO with a KL-divergence coefficient $\beta = 0.01$. We generate $G = 4$ outputs for each prompt to compute group relative advantages. All experiments were conducted on 4 $\times$ NVIDIA A800 GPUs.

5.2 Baselines

To comprehensively evaluate the effectiveness of **VIDA**, we compare it against three categories of state-of-the-art models:

General VLMs. We select a comprehensive set of open-source VLMs to benchmark general capabilities. This includes the classic LLaVA-1.5-7B, the widely-used DeepSeek-VL-7B, and the state-of-the-art InternVL series (3.5-8B and 3-38B) and Qwen series (2.5-VL and 3-VL). This selection covers diverse architectures and scales, testing whether our 8B model can outperform larger (38B) or newer general-purpose baselines.

Text-only LLMs. We employ Llama-3-8B and DeepSeek-V3 to represent the "logic ceiling" of pure text models. These baselines decouple linguistic reasoning from visual perception, verifying whether visual processing is essential for resolving ambiguous design intents.

SFT Baseline. Qwen3-VL-8B + SFT is our backbone model trained with supervised fine-tuning.

5.3 Main Results

The comparative results of clarification generation are presented in Table 3. Our proposed **VIDA** framework significantly outperforms all baselines across key strategic and accuracy metrics.

Superiority in Strategic Planning. **VIDA** achieves a Strategic Alignment Score (SAS) of 49.02%, surpassing the SFT baseline by +20.59%. This significant margin confirms that our GSPO training effectively instills the expert priority protocol ($L_3 \succ L_1$), whereas SFT merely learns fluent but strategy-agnostic patterns.

Accuracy against Logic Giants. With a Keyword Recall (K-RME) of 82.35%, **VIDA** outperforms the logic-heavy DeepSeek-V3 (73.53%). This result highlights the limitation of caption-based reasoning in complex design tasks and validates the necessity of **VIDA**'s end-to-end visual perception for capturing subtle missing elements.

Qualitative Analysis. As visualized in Figure 3, **VIDA** demonstrates superior semantic alignment with expert designers (highlighted in red). Unlike baselines that generate generic inquiries or get distracted by irrelevant objects (e.g., focusing on "attire" rather than "room style"), **VIDA** successfully grounds abstract user needs into concrete design elements and proactively identifies specific visual defects in renovation contexts.

5.4 Ablation Study

We perform ablations on VIDA to isolate the contribution of each reward term, the RL algorithm, and the two-stage training pipeline. Results are shown in Table 4.

Model Variant	SAS (↑)	V-G (↑)	K-RME (↑)
VIDA (Full)	**49.02**	**0.5035**	**82.35**
Reward Ablation			
w/o R_{Hier}	31.25	0.4891	76.18
w/o R_{Vis}	47.83	0.3124	79.42
w/o R_{Sim}	46.91	0.4912	74.56
w/o R_{Fmt}	48.15	0.4987	81.03
Algorithm Ablation			
GSPO → PPO	45.67	0.4823	79.88
GSPO → GRPO	47.21	0.4901	80.45
Training Stage Ablation			
RL-from-scratch	38.94	0.3567	71.23
SFT-only	28.43	0.2970	54.90

Table 4: Ablation study of VIDA on the validation set. Best results are shown in bold.

Effect of reward components. Removing R_{Hier} causes the largest drop in SAS, showing that hierarchical supervision is the key source of VIDA's priority-aware questioning ability. Removing R_{Vis} sharply degrades visual grounding, confirming its role in mitigating modality forgetting during RL. Removing R_{Sim} mainly reduces K-RME, suggesting that it helps preserve semantic alignment with expert questions. By contrast, R_{Fmt} has only a minor effect.

Effect of policy optimization algorithm. Replacing GSPO with PPO or GRPO consistently reduces performance, indicating that group-relative optimization is better suited to our composite reward design.

Effect of training strategy. The full two-stage pipeline is also necessary. RL-from-scratch underperforms the full model on all metrics, while SFT-only yields the worst SAS, showing that SFT alone cannot induce strategic reasoning and RL alone cannot provide stable multimodal grounding.

Cross-room-type generalization. We further perform leave-one-room-type-out evaluation to test whether VIDA overfits to room-specific templates. Compared with the in-domain SAS of 49.02%, VIDA achieves 45.81%, 44.23%, 45.67%, and 46.83% when testing on bathrooms, kitchens, bedrooms, and living rooms, respectively. The limited performance drop suggests that VIDA captures transferable priority-aware reasoning within the interior design domain rather than simply memorizing room-specific patterns.

Overall, these results show that VIDA's gains stem not only from its component design, but also from its ability to generalize across room-type variations.

5.5 Expert Preference Study

To further evaluate the practical utility of VIDA in real-world interior design consultations, we conduct a blind expert preference study. We randomly sample 30 cases from the test set and present the

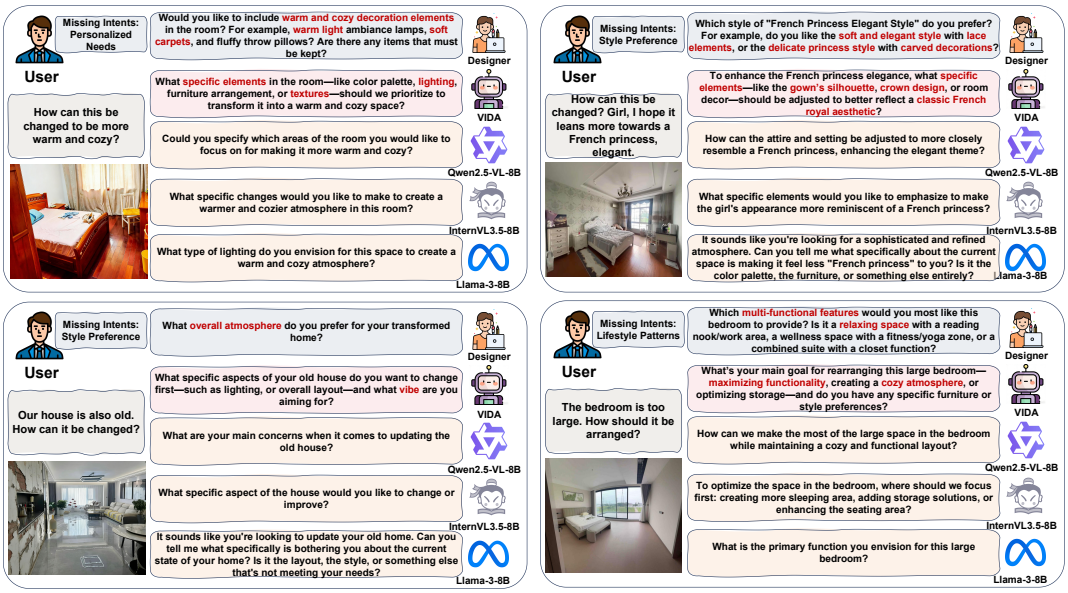

Figure 3: **Qualitative Comparison.** Red highlights indicate semantic alignment between the expert (ground truth) and **VIDA**, suggesting that **VIDA** captures key strategic intents. Baselines often fail by overfocusing on salient image subjects (top-right, attire over room style) or treating the task as repair-oriented rather than eliciting ambience and design vision (bottom-left).

clarification questions generated by VIDA and the baseline side by side, with randomized order and hidden model identities. Evaluators with real-world interior design consultation experience choose among "VIDA", "Baseline" and "Tie" based on which question better supports subsequent design communication and clarification.

The results show an overall preference for VIDA. Excluding ties, VIDA achieves a preference rate of 73.2%, compared with 26.8% for the baseline. This suggests that VIDA not only performs better on automatic metrics, but also generates clarification questions that are more helpful in realistic design consultation settings.

6 Discussion

In this section, we analyze the cognitive shifts enabled by our method, focusing on the transition to value-driven reasoning and the role of visual perception in strategy formulation.

Transition to Value-Driven Strategic Reasoning. Standard VLMs operate on likelihood maximization, often favoring generic, safe queries to minimize perplexity. In contrast, **VIDA** introduces a *value-driven* decision mechanism. By internalizing hierarchical rewards, the model learns that resolving high-level ambiguities yields higher strategic value than addressing trivial details. This enables *structured top-down reasoning*, where the assistant establishes design constraints before variables, effectively mirroring professional cognitive protocols.

Visual-Strategic Interdependence. Our analysis reveals that **VIDA** utilizes visual cues as condition variables to drive strategic branching. For instance, the model maps cluttered scenes to storage optimization pathways, while empty spaces trigger spatial functionality inquiries. The observed "Visual Recovery" thus represents a strategy-aware visual attention mechanism, where the assistant actively attends to specific image regions solely to support the formulation of high-priority questions.

7 Conclusion

We introduced **VIDA**, a Visual Intent-driven Design Assistant designed to bridge the gap between general multimodal understanding and professional consultation strategies. By proposing a two-stage training framework, we successfully aligned the model with hierarchical design protocols. Crucially, our novel strategic reward mechanism, featuring Dynamic Intent Binding, enables the assistant to prioritize high-level style inquiries while maintaining semantic consistency. Experiments show that **VIDA** consistently outperforms strong VLM and LLM baselines in strategic alignment. A blind expert preference study further indicates that **VIDA** is preferred for clarification in realistic interior design consultation settings. Future work aims to generalize our hierarchical reward to multi-step reasoning and extend the framework to other professional consultation domains.

Acknowledgements

This project is partially supported by Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things (No. 2023B1212010007), the CNPC Technology Project "Research on Key Technologies of Artificial Intelligence for Oil and Gas Exploration and Development" (2023DJ84), China NSFC Grant (62472366), the Project of DEGP (No. 2023KCXTD042, 2024GCZX003), the "111" Center (No. D25008), and the Shenzhen Science and Technology Foundation (ZDSYS20190902092853047).

Limitations

Our work currently has limitations spanning technical optimization and data diversity.

From a technical perspective, although visual grounding is improved, the model still struggles with fine-grained spatial reasoning in highly occluded scenes, occasionally misinterpreting depth relationships in cluttered rooms. Furthermore, we observe a trade-off between strategic alignment and linguistic diversity. The GSPO optimization, driven by strong hierarchical rewards, occasionally leads to strategy overfitting, where the assistant converges on a limited set of high-reward question templates to maximize scores, potentially reducing the naturalness and variability of expression.

From a data perspective, **InteriorClarify** primarily consists of consultations from Chinese social platforms, which may limit the model's generalizability to design norms and housing structures in other cultural contexts. Additionally, a subset of instances contains synthetic images generated from text. While these images maintain visual-textual consistency, they may not fully capture the complex lighting, noise, and "messy" details of real-world photography, potentially affecting the model's robustness when deployed in unrestricted environments.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736.

Mohammad Aliannejadi, Julia Gisle, Aleksandr Chuklin, Jeff Dalton, and Mikhail Burtsev. 2021. Building and evaluating open-domain dialogue corpora with clarifying questions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4473–4484, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. 2019. Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 475–484, Paris, France. ACM.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Yue Feng, Hossein A. Rahmani, Aldo Lipani, and Emine Yilmaz. 2023. Towards asking clarification questions for information seeking on task-oriented dialogues. *Preprint*, arXiv:2305.13690.

Menghong Guan, Subrota Kumar Mondal, Hong-Ning Dai, and Haiyong Bao. 2023. Reinforcement learning-driven deep question generation with rich semantics. *Information Processing and Management*, 60(2).

Xiang Hu, Zuijie Wen, Yafang Wang, Xiaolong Li, and Gerard De Melo. 2020. Interactive question clarification in dialogue via reinforcement learning. In *Proceedings of the 28th International Conference on Computational Linguistics: Industry Track*, pages 78–89, Online. International Committee on Computational Linguistics.

Pu Jian, Donglei Yu, Wen Yang, Shuo Ren, and Jiajun Zhang. 2025. Teaching vision-language models to ask: Resolving ambiguity in visual questions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3619–3638. Association for Computational Linguistics.

Kasia Kobalczyk, Nicolás Astorga, Tennison Liu, and Mihaela van der Schaar. 2025. Active task disambiguation with LLMs. In *The Thirteenth International Conference on Learning Representations*. ICLR 2025 Spotlight.

Vaibhav Kumar and Alan W Black. 2020. Clarq: A large-scale and diverse dataset for clarification question generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7296–7301, Online. Association for Computational Linguistics.

Wai-Chung Kwan, Hong-ru Wang, Huimin Wang, and Kam-Fai Wong. 2023. A survey on recent advances and challenges in reinforcement learning methods

for task-oriented dialogue policy learning. *Machine Intelligence Research*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pages 12888–12900.

Ling Li, Yao Zhou, Yuxuan Liang, Fugee Tsung, and Jiaheng Wei. 2025. Recognition through reasoning: Reinforcing image geo-localization with large vision-language models. *arXiv preprint arXiv:2506.14674*.

Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. 2023. We're afraid language models aren't modeling ambiguity. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 790–807.

Brielen Madureira and David Schlangen. 2023. Instruction clarification requests in multimodal collaborative dialogue games: Tasks, and an analysis of the codraw dataset. *Preprint*, arXiv:2302.14406.

Ninareh Mehrabi, Mohammad Alouf, Apurv Sharma, Anurag Rao, and Ahmad Beirami. 2023. Resolving ambiguities in text-to-image generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7051–7074.

Fangwen Mu, Lin Shi, Song Wang, Zhuohao Yu, Binquan Zhang, Chenxue Wang, Shengze Liu, and Qing Wang. 2024. Clarifygpt: A framework for enhancing llm-based code generation via requirements clarification. In *Proceedings of the 2024 ACM International Conference on the Foundations of Software Engineering (FSE)*, pages 89–101.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Boyuan Pan, Hao Li, Ziyu Yao, Deng Cai, and Huan Sun. 2019. Reinforced dynamic reasoning for conversational question generation. *Preprint*, arXiv:1907.12667.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

Kimia Ramezan, Alireza Amiri Bavandpour, Yifei Yuan, Clemencia Siro, and Mohammad Aliannejadi. 2025. Multi-turn multi-modal question clarification for enhanced conversational understanding. *Preprint*, arXiv:2502.11442.

Sudha Rao and Hal Daume III. 2018. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2737–2746, Melbourne, Australia. Association for Computational Linguistics.

Sudha Rao and Hal Daumé. 2019. Answer-based adversarial training for generating clarification questions. *Preprint*, arXiv:1904.02281.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Zhengxiang Shi, Yue Feng, and Aldo Lipani. 2022. Learning to execute actions or ask clarification questions. *Preprint*, arXiv:2204.08373.

Zhengxiang Shi, Jerome Ramos, To Eun Kim, Xi Wang, Hossein A. Rahmani, and Aldo Lipani. 2023. When and what to ask through world states and text instructions: Iglu nlp challenge solution. *Preprint*, arXiv:2305.05754.

Han Sun, Qin Li, Peixin Wang, and Min Zhang. 2026. Mitigating object hallucinations in lvlms via attention imbalance rectification. *arXiv preprint arXiv:2603.24058*.

Zhen Sun, Ziyi Zhang, Zeren Luo, Zhiyuan Zhong, Zeyang Sha, Tianshuo Cong, Zheng Li, Shiwen Cui, Weiqiang Wang, Jiaheng Wei, and 1 others. 2025. Can vlms detect and localize fine-grained ai-edited images? *arXiv preprint arXiv:2505.15644*.

Shuai Wang, Zhenhua Liu, Jiaheng Wei, Xuanwu Yin, Dong Li, and Emad Barsoum. 2025. Athena: Enhancing multimodal reasoning with data-efficient process reward models. *arXiv preprint arXiv:2506.09532*.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, and 20 others. 2025. Qwen-image technical report. *Preprint*, arXiv:2508.02324.

Jingjing Xu, Yuechen Wang, Duyu Tang, Nan Duan, and Pengcheng Yang. 2019. Asking clarification questions in knowledge-based question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1618–1629, Hong Kong, China. Association for Computational Linguistics.

Yifei Yuan, Clemencia Siro, Mohammad Aliannejadi, Maarten De Rijke, and Wai Lam. 2024. Asking multimodal clarifying questions in mixed-initiative conversational search. In *Proceedings of the ACM Web Conference 2024*, pages 1474–1485, Singapore, Singapore. ACM.

Hamed Zamani, Gord Lueck, Everest Chen, Rodolfo Quispe, Flint Luu, and Nick Craswell. 2020. Mimics: A large-scale data collection for search clarification. *Preprint*, arXiv:2006.10174.

J. D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why johnny can't prompt: How non-ai experts try (and fail) to design llm prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

He Zhang, Ying Sun, Weiyu Guo, Yafei Liu, Haonan Lu, Xiaodong Lin, and Hui Xiong. 2023. Interactive interior design recommendation via coarse-to-fine multimodal reinforcement learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6472–6480. ACM.

Michael J. Q. Zhang, W. Bradley Knox, and Eunsol Choi. 2025a. Modeling future conversation turns to teach llms to ask clarifying questions. *Preprint*, arXiv:2410.13788.

Ziyi Zhang, Zhen Sun, Zongmin Zhang, Zifan Peng, Yuemeng Zhao, Zichun Wang, Zeren Luo, Ruiting Zuo, and Xinlei He. 2025b. "I Can See Forever!": Evaluating Real-time VideoLLMs for Assisting Individuals with Visual Impairments. *arXiv preprint arXiv:2505.04488*.